MULTIMEDIA UNIVERSITY

STUDENT ID NO

MULTIMEDIA UNIVERSITY

FINAL EXAMINATION

TRIMESTER 2, 2017/2018

TDS2101 – INTRODUCTION TO DATA SCIENCE

(All sections / Groups)

12th MARCH 2018
9.00 AM – 11.00 AM
(2 Hours)

INSTRUCTIONS TO STUDENTS

1. This Question paper consists of 4 pages with 4 Questions only excluding the cover page.

2. Attempt **ALL** questions. All questions carry equal marks and the distribution of the marks for each question is given.

3. Please print all your answers in the Answer Booklet provided.

QUESTION 1

a) Describe FOUR differences between *business intelligence* and *data science.* (4 marks)

b) Differentiate between systems that support *business intelligence* and systems that support *data science with analytics* by providing an example for each type of system as mentioned above. (3 marks)

c) Describe THREE responsibilities of a data scientist. (3 marks)

CONTINUED...

QUESTION 2

a) The following table contains two columns namely a variable *Age* with continuous values and a class label / class variable *PlayBall.*

Age	*PlayBall*
40	Yes
43	Yes
44	Yes
59	Yes
62	Yes
62	Yes
65	Yes
67	Yes
50	No
54	No
54	No
55	No
55	No
57	No
57	No
57	No
63	No
67	No
67	No
68	No

(i) Discretize the values of variable *Age* using entropy-based binning method for the following intervals. Calculate the Gain for each interval. Round off your answer to two decimal places.

1. (Age <=50, Age > 50) (2 marks)
2. (Age <=60, Age > 60) (2 marks)

(ii) Based on your answers in Question 2(a)(i), which is the best interval (1. or 2.)? Justify your answer. (2 marks)

b) A client for a project suggests MongoDB as its new database for storing data. It is a NoSQL database.

(i) There are four types of NoSQL databases. State the type of database for MongoDB. (1 mark)

(ii) Provide an example of how the architecture of MongoDB has an advantage over relational database management systems. (2 marks)

c) State the name of code repositories that contain R packages and libraries for Python. (1 mark)

CONTINUED...

QUESTION 3

a) What is the purpose of partitioning a set of data into a training set and a test set? (2 marks)

b) Given below is a list of transactions that occur in a stationery shop between 9.00am – 9.30am on Sunday.

Transaction ID	Item(s)
1	Pen, Stapler, Scissors , Paper
2	Pencil, Scissors, Paper, Pen
3	Envelope, Stapler
4	Pen, Pencil
5	Paper, Envelope, Pen, Pencil
6	Scissors , Pencil, Paper, Envelope,
7	Pen, Paper, Stapler
8	Pencil, Stapler, Envelope
9	Pen, Envelope, Paper, Scissors
10	Pen
11	Scissors, Pen, Paper,

(i) Given the following two association rules, calculate the confidence and lift for the rules. For confidence values, provide your solution in percentage. Round off your answer to two decimal places. (4 marks)

1. Paper → Scissors
2. Pen → Pencil

(ii) Based on the confidence and lift values of two association rules, advise the stationery shop owner on layout arrangement for paper, pen, pencil and scissors. (1 mark)

c) The table below shows the ratings for movies given by anonymous users, ranging from 0 to 1. Apply Euclidean distance and round off your solution to two decimal places. Based on your solution, identify which movies can be grouped and not grouped together. (3 marks)

	User1	User2	User3	User4	User5	User6
MovieX	0.6	0.9	0.8	0.9	0.5	0.7
MovieY	0.5	0.8	0.6	0.4	0.5	0.2
MovieZ	0.7	0.8	0.8	0.9	0.6	0.7

CONTINUED...

QUESTION 4

Figure 1 (as below) shows the content of employee records in a comma separated value file named *TDS2101.csv*. For Question 4(a) and Question 4(b), refer to Figure 1.

Figure 1: TDS2101.csv

```
id,name,position,allowance,status,location
100,Gloria,clerk,2000,married,cyberjaya
200,Ahmad,lecturer,5000,single,sekinchan
300,Jimmy,manager,4500,single,puchong
400,Ragu,security,2500,married,banting
500,Simon,clerk,2300,married,sepang
600,Lydia,secretary,3000,single,puchong
700,Michelle,professor,8000,single,melaka
800,Ali,lecturer,5500,married,johor
```

a) Using R, create a function with no argument list and assign it to variable *onefunction.* Assume that R's working directory is set to the right location to retrieve files. This function loads the TDS2101.csv to R's working environment and displays the difference/gap between the maximum and minimum salary of employees. (4 marks)

b) Using R, create a function with no argument list and assign it to variable *twofunction.* Assume that R's working directory is set to the right location to retrieve files. *twofunction* loads the TDS2101.csv to R's working environment. Except for values for variables *id* and *allowance*, the values of other variables should be loaded as factor. Then, this function will replace the value for variable *status* from married to M. Even after the changes have been completed, the structure of the objects remains the same as the initial structure during the loading of the file to R's working environment. (4 marks)

c) Your department head would like to know the number of participants for all the data science courses that you have conducted within a month. Will table or graph be suitable to present this information to your department head? Justify your answer. (2 marks)

END OF PAGE.